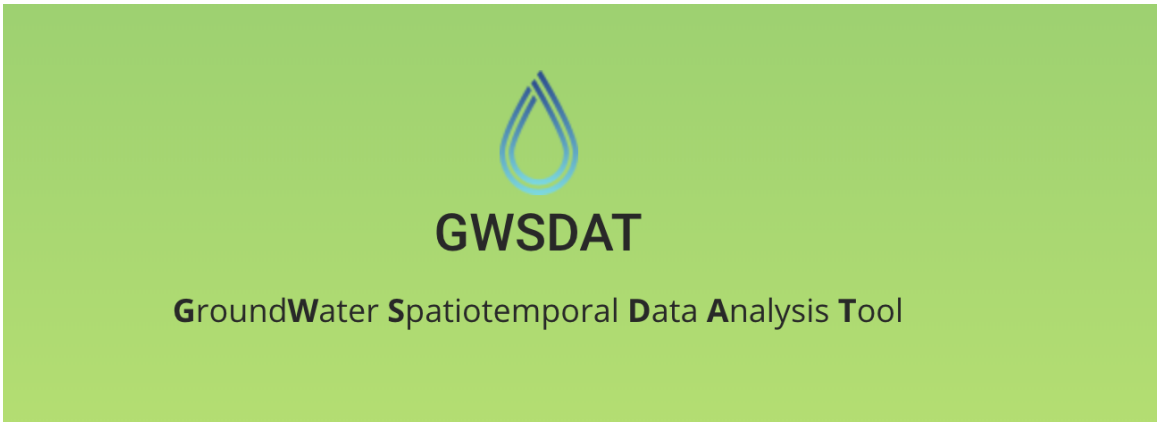

Version 3.2 User Manual

Wayne R. Jones (Wayne.W.Jones@shell.com)
Principal Data Scientist
Shell Research Limited (U.K.) Ltd, London

Luc Rock (Luc.Rock@shell.com)
Soil and Groundwater Scientist
Shell Global Solutions International B.V., Netherlands, Rijswijk

Claire Miller (Claire.Miller@glasgow.ac.uk)
Marnie Low (Marnie.Low@glasgow.ac.uk)
Craig Alexander (Craig.Alexander.2@glasgow.ac.uk)
Adrian Bowman (Adrian.Bowman@glasgow.ac.uk)
School of Mathematics & Statistics, The University of Glasgow

11 January, 2023

Contents

List of Figures

1 Acknowledgements

The authors gratefully acknowledge the many different people who have willingly contributed their knowledge and their time to the development of GWSDAT[1].

The authors wish to express their gratitude to Ludger Evers and Daniel Molinari from the Department of Statistics, University of Glasgow, for their invaluable contributions to the statistical aspects of GWSDAT. Thanks also to Ewan Crawford for his assistance in the development of the original GWSDAT user interface.

We acknowledge and thank the R project for Statistical Computing and all its contributors without which this project would not have been possible.

A big thank you to Shell's worldwide environmental consultants for assistance in evaluating and testing GWSDAT. Thanks also to the Shell Year in Industry students Tess Brina, Rosemary Archard, Emma Toms, Stephanie Marrs and Rachel Stroud who spent a great deal of time using GWSDAT and making suggestions for improvements.

We thank our colleagues Matthew Lahvis, George Devaull, Matthijs Bonte, Hayley Thomas, Karina Cady, Jonathan Smith from the Shell SGW Projects & Technology and FDG teams and Philip Jonathan, Shell Chief Statistician, for their support, vision and advocacy of GWSDAT.

The original idea of GWSDAT was inspired by Marco Giannitrapani.

2 Introduction

The GroundWater Spatiotemporal Data Analysis Tool (GWSDAT[2]) has been developed by Shell Global Solutions and the University of Glasgow to help visualise trends in groundwater monitoring data. It is designed to work with simple time-series data for solute concentration and ground water elevation, but can also plot non-aqueous phase liquid (NAPL) thickness if required. Spatial data is input in the form of well coordinates, and wells can be grouped to separate data from different aquifer units. The software also allows the import of a site basemap in GIS shapefile format. Trend and contour plots generated using GWSDAT can be easily exported directly to a range of different formats, including Microsoft PowerPoint.

The underlying geostatistical calculations are generated using the open source statistical program R (R Development Core Team (2008)). Since version 3.0 the graphical user interface is the open source web framework R Shiny package (`https://shiny.rstudio.com`) which enables both local and online deployment. More details on the statistical methods can be found in the Appendices in Section 7.

Potential applications where GWSDAT can add value (cost savings and reduction in environmental liabilities) through improved risk-based decision making and response include:

- Early identification of increasing trends or off-site migration.
- Evaluation of groundwater monitoring trends over time and space (i.e., holistic plume evaluation).
- Nonparametric statistical and uncertainty analyses to assess highly variable groundwater data.
- Reduction in the number of sites in long-term monitoring or active remediation through simple, visual demonstrations of groundwater data and trends.
- More efficient evaluation and reporting of groundwater monitoring trends via simple, standardised plots and tables created at the 'click of a mouse'.

[2]**Disclaimer**: There is no warranty for the Program (GWSDAT), to the extent permitted by applicable law. SHELL, Affiliates of SHELL, the copyright holders and/or any other party provide the Program 'as is' without warranty of any kind, either expressed or implied, including, but not limited to, the implied warranties of merchantability and fitness for a particular purpose. The entire risk as to the quality and performance of the Program is with the LICENSEE. Should the Program prove defective, the LICENSEE assumes the cost of all necessary servicing, repair or correction.

3 Accessing GWSDAT

This section describes the three different ways for users to gain access to GWSDAT. Please refer to Section 6: GWSDAT Graphical User Interface, for details on how to use the tool.

3.1 Online Version

With no software installation required, the easiest way is to use the online version, available at `www.gwsdat.net`. This web site also has general information about the software tool, including help files, videos and case studies (`www.gwsdat.net/case-studies`). The underlying architecture is the GWSDAT `R package` deployed as an app on a `Shiny server` hosted by the University of Glasgow's School of Mathematics and Statistics. Note that the data the user enters is sent over the internet and analysed on this server. See the GWSDAT GitHub development page `www.github.com/WayneGitShell/GWSDAT` for more details on how to deploy on a Shiny server.

3.2 Excel Add-in Interface

The most traditional and widely used method to access GWSDAT is via the Excel Add-in interface. The latest version together with installation instructions and supporting information can be found on the industry websites:

- American Petroleum Institute (API): `www.api.org/GWSDAT`
- Contaminated Land: Applications in Real Environments (CL:AIRE): `www.claire.co.uk/GWSDAT`

This method of deployment involves installing the open source statistical program R (R Development Core Team (2008)) and Excel Add-in locally on a user's computer. All data sets are retained and analysed locally with no information being sent over the internet. Please note that administrator rights may be required for successful installation. Please refer to Section5: GWSDAT Excel Add-in more details.

3.3 GWSDAT R Package

GWSDAT uses the widely available, open source, statistical program `R` (R Development Core Team 2008). This should be downloaded from `www.r-project.org` where versions for all major computing platforms are available. The user may also find it convenient to install the *RStudio* 'Integrated Development Environment' for `R`, freely available from `www.rstudio.com`. This manages some aspects of the `R` environment in a helpful way. When `R` or `RStudio` is launched, one of the visible windows is a 'console'. GWSDAT is available as a package in `R` and this can be installed by typing the instruction

```
install.packages("GWSDAT")
```

in the console window. Note the capital letters, as `R` is case-sensitive. The package is retrieved from the `R` archive `CRAN`, so an internet connection is required for this step. The package will then be installed locally. GWSDAT uses several other `R` packages and these will be installed at the same time. There may be a warning message about a mismatch between the version of `R` used to build the package and the version of `R` installed on your computer if this is not the most recent one, but this is unlikely to cause any difficulty. This local installation step is required only once. The benefit of local installation is that GWSDAT will now be available at any time, with or without an internet connection.

To launch GWSDAT, issue the following two instructions in the console window:

```
library(GWSDAT)
launchApp()
```

The first instruction loads the package so that it can be used in the current session of `R`. The second instruction launches GWSDAT.

See `www.github.com/WayneGitShell/GWSDAT` for the latest development version together with more advanced information and details on how to use the GWSDAT R package.

4 Input Data Format

There are a few different ways to enter user data into GWSDAT and the available choices will depend upon the way the user is accessing GWSDAT. However, no matter, which way you are Accessing GWSDAT the formatting standard and conventions of the raw data input tables are identical. Hence, this section is dedicated to describing the standardised data table input structure and formatting conventions. Later sections will discuss the data input details with respect to the mode of access. See Section 5 for data input via Excel Add-in and Section 6.1 for data input directly via the Graphical User Interface (in case of tool installation as per Section 3.1 and 3.3).

There are only two input tables that must be completed, namely, the *Historical Monitoring Data* table and the *Well Coordinates* table. Please note that there should be no empty rows in either table. Optionally, users can also specify links to the location of GIS shapefiles for use as basemaps or site plans.

4.1 Historical Monitoring Data Input Table

Each row of this table corresponds to a unique combination of well, sampling date and solute type. Groundwater and NAPL (Non-Aqueous Phase Liquids) gauging data may also be entered in this table. Figure 1 displays an example GWSDAT input data set for illustrative purposes. The columns (fields) in the *Historical Monitoring Data* input table are as follows:

Historical Monitoring Data

WellName	Constituent	SampleDate	Result	Units	Flags
SGS5 P1	Nitrate	05/11/2009	54.59	mg/l	E-acc
SGS5 P2	Nitrate	05/11/2009	67.93	mg/l	
SGS5 P1	Sulphate	05/11/2009	99	mg/l	E-acc
SGS5 P2	Sulphate	05/11/2009	61	mg/l	
GDBH102	Ethylbenzene	03/11/2009	ND<1	ug/l	
GDBH104	Ethylbenzene	03/11/2009	ND<1	ug/l	
GDBH104	Toluene	03/11/2009	ND<1	ug/l	
GDBH104	TPH	03/11/2009	36	ug/l	
MW10	Ethylbenzene	03/11/2009	ND<1	ug/l	
MW10	Toluene	03/11/2009	ND<1	ug/l	
MW10	TPH	03/11/2009	ND<5	ug/l	
MW103	Ethylbenzene	03/11/2009	9	ug/l	
MW103	Toluene	03/11/2009	6	ug/l	
MW103	TPH	03/11/2009	162	ug/l	

Figure 1: Example Historical Monitoring table

- **WellName**: the name or identifier of the well (or soil boring) from which the sample was collected. Well names must be consistent and unique. For example, 'MW-1' and 'MW1' will be treated as different wells.
- **Constituent**: Here enter the name of the solute type, e.g. Benzene, Toluene. Again in the same manner as WellName please ensure that the name of a solute is consistent and unique for all entries. The identifiers 'GW' and 'NAPL' are reserved for Groundwater elevation measurements and NAPL thickness data respectively, see further details below.
- **SampleDate**: the date at which the well was sampled (not the date the results were returned from laboratory analysis). Please use a calendar date format, the preferred format is 'dd/mm/yyyy'. Do not include a time of day.
- **Result**: the value of the measurement made. This will be a solute concentration, a groundwater level or a NAPL thickness, as specified in the *Constituent* column.
  - **Solute Concentrations**: The concentration of the constituent is entered here. Non-detect values should be entered as either '<X' or 'ND<X', where 'X' is the detection limit specified by the

laboratory. For example, if the detection limit is 100ug/l then either '<100' or 'ND<100' is acceptable. The non-detect threshold value must be specified so 'ND' on its own is not permissible. In the absence of known detection limits, a sensible value must be substituted. This could be the lowest measured value for the solute in the dataset.

  - **Groundwater** level data is entered as an elevation above a common datum, such as metres or feet above sea level or some other common reference height. All groundwater measurement entries should have the same units, such as metres or feet, and the 'Constituent' field should be set to 'GW'. In the presence of NAPL, please ensure that the groundwater level has been corrected for NAPL density.
  - **NAPL** thickness data is also entered here. Please ensure that all NAPL thickness entries have the same units, e.g. feet or metres and that the Constituent field is set to 'NAPL'. If no NAPL is present, do not add a NAPL entry with zero thickness; simply omit from the table. Where NAPL is recorded in soil borings that do not reach the water table the NAPL thickness should be entered as zero. Well location markers for soil borings or wells where NAPL has been recorded are highlighted in red.

- **Units**: Solute concentration data can either be 'mg/l' or 'ug/l'. For groundwater elevation and NAPL thickness data please set to one of 'level' or one of 'mm', 'cm', 'metres', 'inches', 'feet', respectively. Units must be specified for each entry. All entered groundwater elevation measurements must have the same units. Likewise for NAPL thickness.
- **Flags**: Four different flags are available to modify the way in which certain types of data are handled by the software. The 'E-Acc' (Electron Acceptor), 'NotInNAPL' and 'Redox' flags are used to identify input data types which are to be omitted in the event that the user activates the NAPL substitution function (see Section 5.3). Note, that it is only necessary to flag one data row in this way for all rows containing that constituent to be excluded from NAPL substitution. The fourth flag ('Omit') can be used to exclude individual data entry rows from the GWSDAT analysis.

4.2 Well Coordinates Table

The *Well Coordinates* table, see Figure 2, is used to store the coordinates of groundwater monitoring wells or soil borings. For most of the purposes of GWSDAT modelling, it is only the relative distances between wells which are important. This means any arbitrary cartesian coordinate system can be used as long as well coordinate values have an aspect ratio very close to 1, i.e. a unit in the x-coordinate is the same distance as a unit in the y-coordinate. Hence, well coordinates can be measured directly from a map, or given in easting and northing, etc.

Well Coordinates

WellName	XCoord	YCoord	Aquifer	CoordUnits
BH1	551689.43	224468.03		metres
BH2	551679.43	224426.03		
BH3	551661.93	224430.99		
GDBH101	551674.93	224439.03		
GDBH102	551678.76	224469.31		
GDBH103	551696.31	224474.70		
GDBH103A	551696.31	224474.90		
GDBH104	551664.76	224459.11		
GDBH105	551696.93	224459.03		
MW1	551695.93	224452.53		
MW10	551636.84	224440.36		

Figure 2: Example Well Coordinates table.

- **WellName**: the name or identifier of the well or soil boring. Well names must be identical to those specified in the *Historical Monitoring Data* table. On a point of detail, it is better to name wells using the convention of 'MW-01' rather than 'MW1' so that plots in GWSDAT are correctly ordered.
- **XCoord**: the x-coordinate of the well.
- **YCoord**: the y-coordinate of the well.
- **Aquifer**: The (optional) aquifer field allows the user to associate wells or soil borings with particular subsurface features (e.g. aquifers, sub-strata), in the event that data from these needs to be modelled separately. The user can enter the name (maximum of 8 characters) of the aquifer or sub-stratum, or select a letter A-G from the drop-down listbox. The aquifer field can also be used to partition the dataset from a large site, in the event that multiple unrelated plumes are present or if wells are clustered with large gaps in between. On initiation of a GWSDAT analysis the user is asked to select an aquifer (subsurface feature) to analyse. **Note**: Plots generated using data associated with particular subsurface features have the feature name appended to the title, e.g. Shallow aquifer. If the user leaves the aquifer flag as blank, no such appending will occur.
- **CoordUnits**: Either leave this field blank or select 'metres' or 'feet'. The units specified in this field are used in the calculation of plume mass balance parameters (e.g. plume area and solute mass), for further details see Section 6.3 on plume diagnostics.

4.3 GIS ShapeFiles

A site plan can be superposed over plots of concentration distribution, NAPL thickness and groundwater elevation (see example in Figure 7). Site plans are imported into GWSDAT in the form of shapefiles (see `http://en.wikipedia.org/wiki/Shapefile` for more information). A shapefile is actually a collection of several files, typically created using ARC-GIS. See Section 7.5 for more details on how to generate a shapefile from a CAD drawing using ARC-GIS.

If using the GWSDAT Excel Add-in interface (Section 5.1) then only the location of the main shapefile (file ending with a '.shp' extension) needs to be specified in the GIS Shapefile table, see Figure 3. The associated data files (e.g. .dbf, .sbn, .sbx, .shx) will be picked up automatically,provided they are in the same folder (see example in Figure 4).It is possible to overlay multiple shapefiles up to a maximum of seven.

Alternatively, if using the GWSDAT Graphical User Interface interface (Section 6.1) to input data then all the associated shapefiles (e.g. .shp, .dbf, .sbn, .sbx, .shx) need to be selected and uploaded.

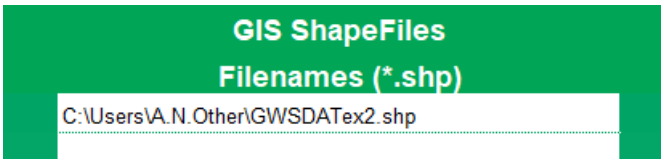

Figure 3: Example GIS ShapeFiles table.

5 GWSDAT Excel Add-in

5.1 Add-in Menu

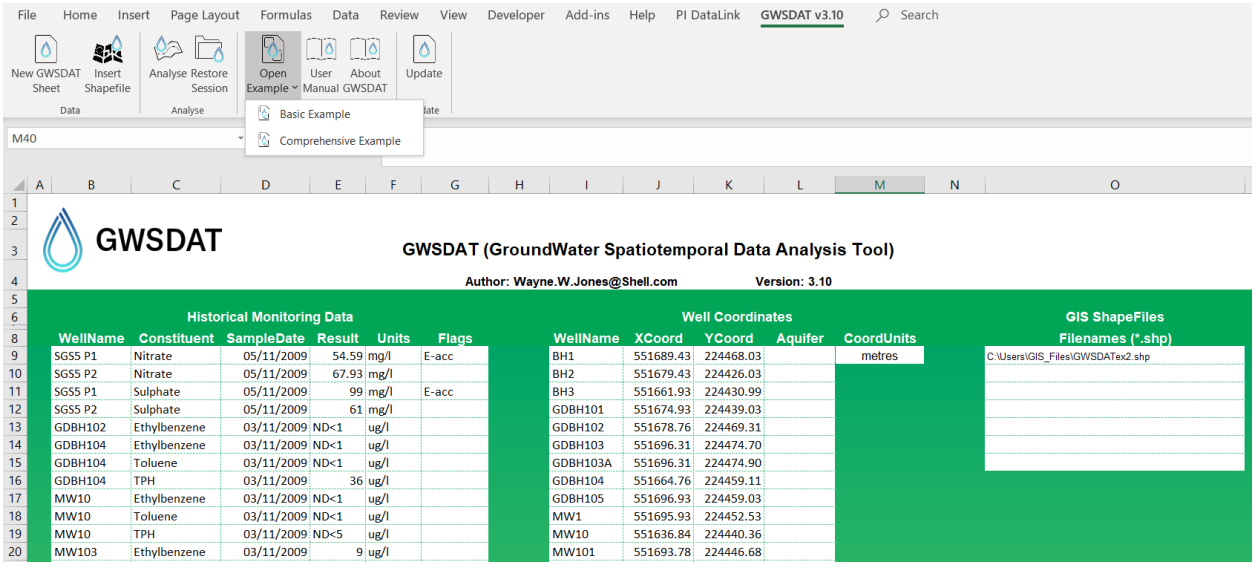

Figure 4: GWSDAT Excel Add-in menu and example data file.

Please see Section 3.2 for details on how to install the GWSDAT Excel Add-in. The menu options, as shown in Figure 4, are as follows:

- **New GWSDAT Sheet**: Inserts a blank GWSDAT data input template worksheet into the active Excel workbook.
- **Insert Shapefile**: Interactively browse for a shapefile and add location to *GIS Shapefiles* table, see Section 4.3 for more information.
- **Analyse**: Begin GWSDAT analysis on the Excel active worksheet data. **Please refer to Sections 6.1 to 6.11 for details on the analysis options and process**.
- **Restore Session**: Load a previously saved GWSDAT session, see Section 6.10 for more information.
- **Open Example -> Basic Example**: Inserts an example GWSDAT worksheet data set into the active Excel workbook.
- **Open Example -> Comprehensive Example**: Inserts a more detailed example GWSDAT worksheet data set which includes a site plan, NAPL thickness data, 'Electron Acceptor' flagged solutes and multiple aquifers into the active Excel workbook.
- **User Manual**: Opens the online GWSDAT user manual via user's web browser. You need be connected to the internet for this to work.
- **About GWSDAT**: Displays version information and Terms & Conditions for GWSDAT.
- **Update GWSDAT**: Check if a newer version of the underlying GWSDAT R package exists at `https://cran.r-project.org/package=GWSDAT`. If a more recent version is detected then it will be automatically installed. You must be connected to the internet for this to work.

5.2 Add-in Data Processing Options

On initiation of a GWSDAT analysis via the Excel Add-in, data processing options are displayed, as shown in Figure 5. The data processing options influence how the data is displayed and how non-detects are handled.

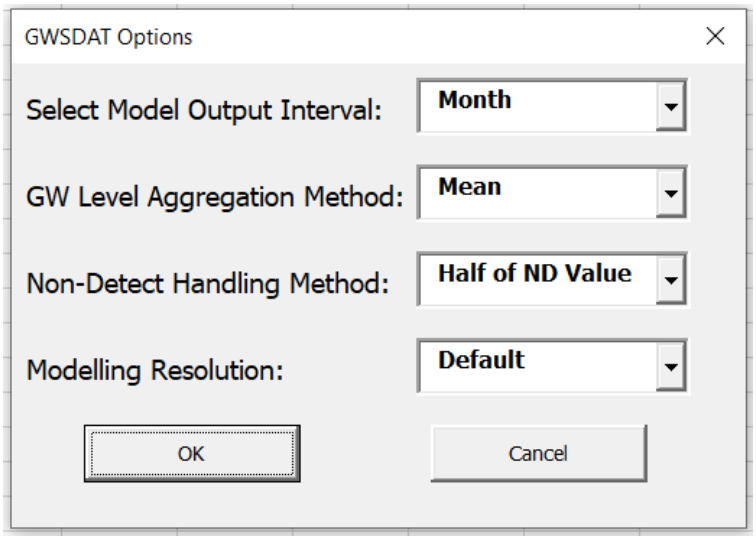

Figure 5: GWSDAT Excel Add-in Processing Options

- **Model Output Interval**: The spatiotemporal model can generate predictions at a user specified interval. The three different options are as follows:
  - **Day**: Concentration and groundwater elevation contour plots are generated for every date represented in the input dataset. This is a good option to choose if each monitoring event comprises samples/ measurements collected within one 24-hour period.
  - **Month**: Concentration and groundwater elevation contours are generated at monthly intervals, working backwards in time from the latest date in the input dataset. Choosing this option aggregates groundwater elevation data within each monthly interval so that a larger dataset is available for the plotting of elevation contours (by local linear regression).
  - **Quarter**: Concentration and groundwater elevation contours are generated at quarterly (3 month) intervals, working backwards in time from the latest date in the input dataset. Choosing this option aggregates groundwater elevation data within each 3-month interval so that a larger dataset is available for the plotting of elevation contours (by local linear regression).

  Note that both the monthly and quarterly model output options only aggregate the dataset used to plot groundwater elevation contours. The solute concentration dataset is not aggregated in time because the spatiotemporal model from which concentration contours are generated does not require this, i.e. the underlying spatiotemporal model used to generate the solute concentration smoother plots does not vary with the data aggregation interval.

  Note that if no monitoring data is present within a particular monthly or quarterly interval, then GWSDAT will not generate a groundwater elevation contour or spatiotemporal solute concentration smoother plot. This is to avoid producing potentially misleading spatial plots far away in time from any actual data.
- **GW Level Aggregation Method**: In the event that there are multiple groundwater elevation measurements from the same well within a given output interval, the user can select how to use this data. The user can select to calculate either the 'Mean', 'Median', 'Min', or 'Max' groundwater elevation. Again, this choice does not affect the spatiotemporal model used to generate the solute concentration smoother plots.
- **Non-Detect Handling Method**: GWSDAT handles non-detect data by a method of substitution. In accordance with general convention, the default option is to substitute the non-detect data with half its detection limit, e.g. ND<50ug/l is substituted with 25ug/l. For a more conservative choice, select

the alternative of non-detect data to be substituted with its full detection limit, e.g. ND<50ug/l is substituted with 50ug/l.

- **Model Resolution**: This option controls the resolution of the spatiotemporal solute concentration smoother (see Appendix 7.1). The user can select between either a default resolution or a higher resolution model fit. In most instances there will be little difference in the modelling results between the two settings. However, in some rare circumstances with complex data sets, it may well be necessary to use the higher resolution setting. Please note it takes approximately 3-4 times longer to fit a higher resolution model.

5.3 NAPL Handling Method

An additional pop-up box will be displayed after the GWSDAT options box if the input contains NAPL data (i.e. 'NAPL' is entered in the constituent field). Selecting 'Yes' to the question 'Do you wish to substitute NAPL values with maximum observed solute concentrations?' forces GWSDAT to recognise NAPL data in the input dataset as indicative of high dissolved solute concentrations. This option has been added to provide the user with a more realistic picture of the area of impacted groundwater in the event that NAPL in wells prevents direct measurement of solute concentrations. *Before using this function the user should, however, be confident that dissolved solutes are derived from the observed NAPL and not from a different source.* Solutes flagged as 'Electron Acceptors' (see Section 4.1) are omitted from the NAPL substitution process.

6 GWSDAT Graphical User Interface (GUI)

The GWSDAT user interface is a web-based graphical user interface (GUI) using the R Shiny framework (`https://shiny.rstudio.com/`) which the user can interact with in many ways. The following sections explain the interface in detail.

6.1 Data Input via Graphical User Interface

If the user entry point to a GWSDAT analysis is via the Excel Add-in interface (Section 5) then the input data has already been specified and the GUI goes straight into *Analyse* mode and is presented as displayed in Figure 7. However, if the user entry point is via the online version or the R package (Sections 3.1 and 3.3)) with no input data specified then a menu is displayed on the left hand side, see Figure 6, initially engaged on the *Manage Data* option. The *Basic Example* and *Comprehensive Example* datasets are all already pre-loaded and are identical to the data sets included in the Excel Add-in interface, see Section 5.1.

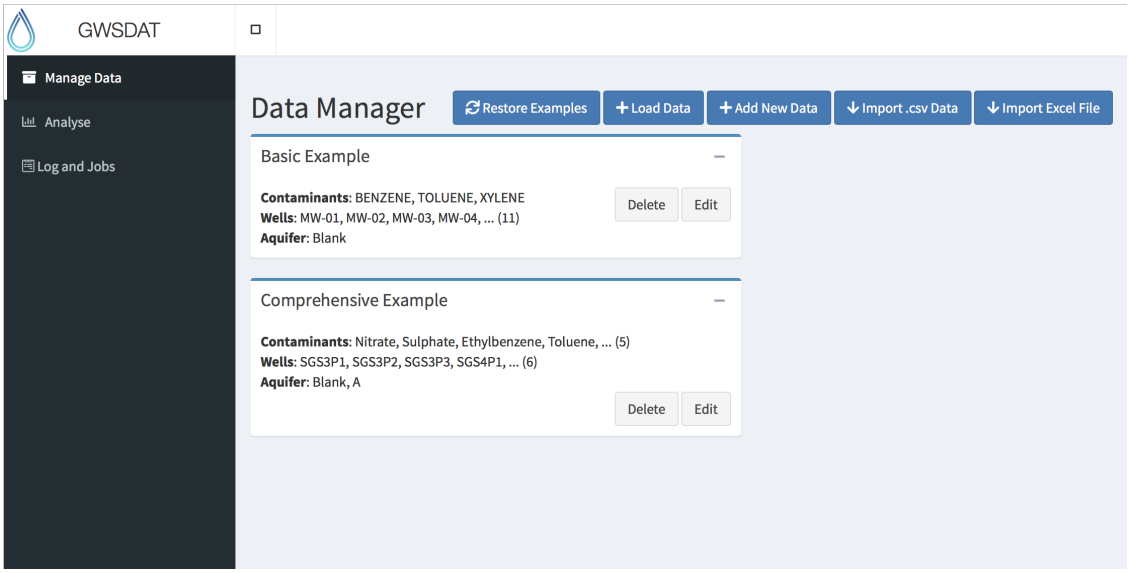

Figure 6: GWSDAT Data Manager.

The Data Manager options allows the user to input data sets in a variety of different formats:

- **Restore Examples**: This allows the two in-built examples to be restored at any point, should this be needed.
- **Load Data**: This allows the user to load a previously saved GWSDAT session, see Section 6.10. Use the *Browse* button to select the '.rds' file, enter an appropriate name and press the *Add Data* button to add the data set to list of available data sets in the Data Manager.
- **Add New Data**: This displays a data editor which allows the user to enter a *Historical Monitoring Data* table and *Well Coordinate data* table in much the same way as data is entered into an Excel spreadsheet. Shapefiles (Section 4.3) can be selected via the *Browse* button. Enter an appropriate name for the data set and click the *Save* button to add the data set to list of available data sets in the Data Manager.
- **Import .csv Data**: This option allows the user to select, via *Browse* buttons, comma separated value (csv) formatted data for the *Historical Monitoring Data* and *Well Coordinate data* tables together with GIS Shapefiles. Example '.csv' formatted input data sets and GIS Shapefiles can be downloaded from `https://github.com/WayneGitShell/GWSDAT/tree/master/data`.
- **Import Excel File**: This options allows the user to input a GWSDAT Excel Add-in formatted input

dataset, see Figure 4, together with GIS Shapefiles. Use the *Browse* button to select the Excel file. If multiple sheets are detected a pop-up menu will prompt the user to select the appropriate sheet. Example input files can be found at `https://github.com/WayneGitShell/GWSDAT/tree/master/inst/extdata`. Enter an appropriate name for the data set and click the *Import* button to add the data set to list of available data sets in the Data Manager.

Please note that data sets can also be specified when launching GWSDAT from the R Statistical Programming Language - see `https://github.com/WayneGitShell/GWSDAT` for details. The user is referred to Section 4 for full details on the format of the input data.

Note: To select a data set for a GWSDAT analysis first select the *Analyse* option on the left hand menu located under the *Manage Data* option.

6.2 Spatial plot

The GWSDAT spatial plot (see Figure 7) is for the analysis of spatial trends in solute concentrations, groundwater flow and, if present, NAPL thickness. It displays the locations of the monitoring wells (black solid dots) together with the well names and actual measured solute concentration values (detect data is displayed in a red font; non-detect in a black font). The date interval for the displayed data is indicated above the spatial plot. If a GIS shapefile has been supplied then the major site features (roads, tanks, etc) are overlaid on the spatial plot as light blue lines.

A key feature of GWSDAT is the ability to produce estimates of contaminant concentrations over space and time simultaneously. This gives a more effective and efficient method of analysis than the examination of concentration maps at isolated time points, or of time trends at isolated locations. The simultaneous use of information over space and time allows estimates at particular locations and times to 'borrow strength' from neighbouring data, see McLean et al. (2019). Use the slider at the foot of the page to explore how the estimates of concentration changes through time. Note that the slider box at the foot of the page can be moved to any convenient position by clicking and dragging with the mouse. The 'Play' symbol (forward-arrow) in the bottom right hand corner of the slider activates a 'movie', which can be paused by pressing the button again. Select a file format (one of `png`, `jpg`, `pdf`, `ps`, `pptx` ,`tif`) from *Image format* and click the *Save plot* button to export individual spatial image plots. The *Generate PPT Animation* function exports the full series of spatial plots directly to Microsoft PowerPoint.

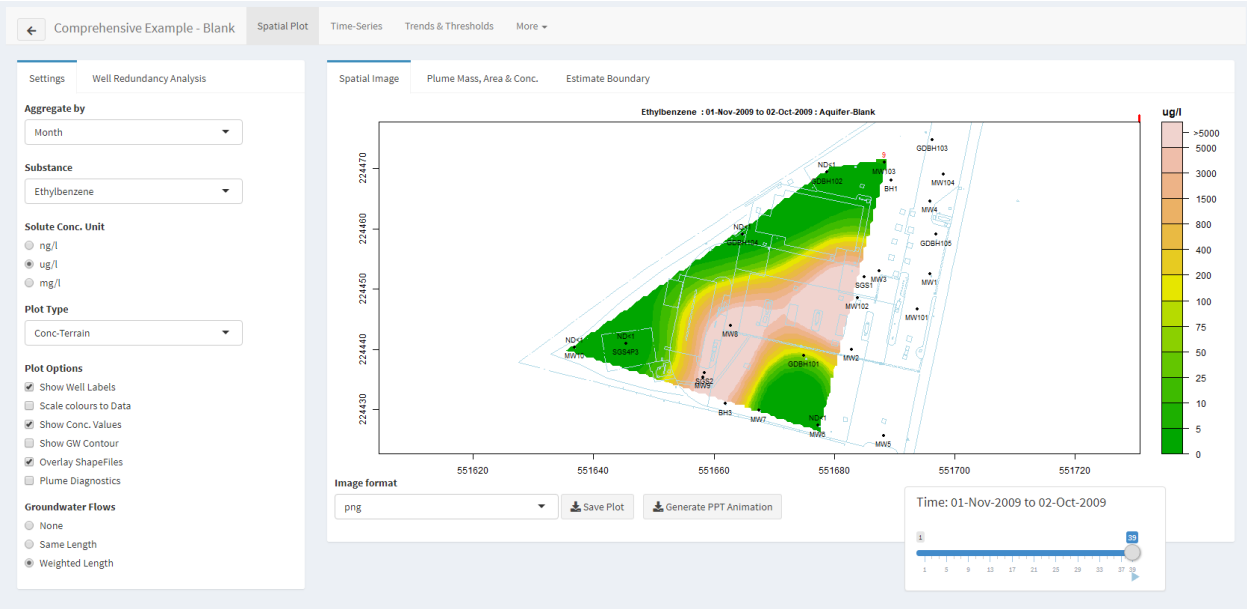

Figure 7: GWSDAT Spatial Plot

The *Settings* tab located to the left hand side of the spatial plot gives control over many aspects of the display:

- **Aggregate by**: provides a drop-down menu which allows the temporal plotting resolution to be altered (Day, Month, Quarter, Year). In the example in Figure 7, a monthly model output interval has been selected and the displayed actual solute concentration values were sampled between the 2nd Oct 2009 and the 1st Nov 2009. This functionality is identical to the 'Model Output Interval' detailed in Section 5.2.
- **Substance**: Drop-down listbox to select the different solutes to be inspected.
- **Solute Conc. Unit**: allows the units to be changed between one of 'ng/l', 'ug/l' or 'mg/l'.
- **Plot Type**: Drop-down listbox with the following choices:
  - **Conc-Terrain**: This option overlays the predictions of the spatiotemporal solute concentration smoother for a particular model output interval using a 'terrain' colour scheme - see example in Figure 7. Please note that the output of the spatiotemporal trend smoother is always given for the latest date in the displayed output interval. The dark green colours indicate low solute concentration and the colours are gradated through yellow and brown to almost white, to illustrate increasing estimated solute concentrations.The concentration values can be read off from the key on the right hand side of the plot. As the user iterates through time steps, it may be noticed that the area covered by the spatiotemporal solute concentration smoother changes. This is due to the fact that spatiotemporal predictions are only generated between interpolated data and are not extrapolated to regions where no data exists, which could potentially lead to erroneous results. For each time step, the area of the contour is calculated only from the collection of wells for which the monitoring period spans the current model output interval. GWSDAT generates predictions in the convex hull region dictated by these wells. The convex hull (see `http://en.wikipedia.org/wiki/Convex_hull`) may be visualised as the expected boundary if an elastic band was placed around the locations of these wells.
  - **Conc-Topo**: This function is identical to Conc-Terrain but uses a topograghic colour scheme which gradates increasing solute concentrations through blue, green, yellow and beige.
  - **Conc-GreyScale**: This function is identical to Conc-Terrain but uses a grey scale colour scheme which gradates increasing solute concentrations through light grey to black. This is useful for printing on black and white printers!
  - **Conc-Terrain-Circles**: This selection overlays (terrain) colour coded circles located at the wells which have been monitored within the current model output interval. The size of the circles scales with the log of the observed solute concentration values and the solute concentration range can be read off from the colour key to the right of the plot.
  - **Conc-Topo-Circles**:This selection is identical to Conc-Terrain-Circles but uses a topographic colour scheme.
  - **Conc-GreyScale-Circles**: This selection is identical to Conc-Terrain-Circles but uses a grey scale colour scheme.
  - **NAPL-Circles**: This selection displays the observed NAPL thicknesses within the current model output interval as size scaled and colour coded circles. NAPL thickness ranges are read off from the colour key on the right hand side of the plot. Colours are gradated from dark red through yellow to almost pure white to illustrate increasing NAPL thickness. The location of wells which have recorded NAPL in any part of their monitoring history are coloured with red solid dots instead of the usual black solid dots.

Hint: In the presence of poor well location network design or limited data then it is recommended the user select either the 'Terrain-Circles' or 'Topo-Circles' plot type.

- **Plot Options**:
  - **Show Well Labels**: This controls whether to display well names/labels immediately below the well locations.
  - **Scale colours to Data**: By default the colour key of solute concentrations is subdivided as shown in Figure 7. By using the same subdivisions the spatiotemporal solute concentration smoother plots can be directly compared between different model output intervals. This control will produce a new colour key whose subdivisions span the concentration predictions for the current model output interval only.
  - **Show Conc. Values**: This controls whether to display actual sampled concentration values immediately above the well locations. If the data is identified as a NAPL measurement the value will be displayed as 'NAPL' in a red font.
  - **Show GW Contour**: To add contour lines of groundwater level data. This superposes isobars of smoothed groundwater elevation data on top of the solute concentration plot. This is achieved through a 2D extension of the local linear regression method described in Appendix 7.4.
  - **Overlay Shapefiles**: This controls whether to overlay a site plan.
  - **Plume Diagnostics**: This controls whether to calculate and display plume diagnostic quantities from the predictions of the spatiotemporal solute concentration smoother (please refer to insert box below and example in Figure 8). The delineated plume is displayed with a solid red contour line which also includes a label displaying the plume boundary threshold value. If the plume boundary lies completely within the area covered by the spatiotemporal solute concentration smoother (i.e. forms a closed loop) then the plume centre of mass is displayed with a red cross and the plume mass and area printed at the bottom left margin of the spatial plot. Note: in order for the correct plume diagnostics units to be used the *CoordUnits* field in the *Well Coordinates* table must be specified, see Section 4.2. More details about plume diagnostics can be found in Section 6.3 and Appendix 7.2.
- **Groundwater Flows**: The blue arrows in Figure 8 display the estimated direction and (relative) hydraulic gradient of groundwater flow at monitoring points across the site. This is calculated from the combination of well coordinates and recorded groundwater elevations for this particular model output interval (see Appendix 7.3 for more details). This radiogroup allows the user to choose either 'not to display groundwater arrows' or 'direction only arrows' or 'both direction and relative strength arrows' (default).

Note: The spatiotemporal solute concentration smoother is a function which simultaneously estimates both the spatial and time series trend in site solute concentrations. By smoothing the data in both space and time it provides a clearer interpretation of site solute concentration dynamics than would otherwise be gleaned from the raw data. However, it is important to note that it is a smoother function and as such, the predictions do not necessarily lie on the observed data points. In the event that a sampled concentration value is significantly larger than the predictions of the spatiotemporal smoother, the well label is coloured red and surrounded by braces, e.g. '<MW-1>'. This serves as a very useful method for outlier detection. In addition, the analysis may be skewed if data are input from monitoring wells with disparate construction or screened in different aquifer systems.
Another important point to consider is that the quality of the spatiotemporal smoother is directly influenced by quality of the underlying data. In general, data originating from sites with many evenly spatially distributed wells with a long time history leads to better quality smoother predictions. The converse of a small number of wells or poor well location network design (e.g. wells located in almost a straight line), or short monitoring history, will lead to less reliable smoother predictions, particularly at the edges.
For the reasons stated above, the predictions should be interpreted with care and a more detailed evaluation may be necessary to understand observed trends and outliers. Further methods for assessing the goodness of fit of the spatiotemporal smoother can be found in Section 6.8. For more details on the spatiotemporal smoothing algorithm, please see Appendix 7.1.

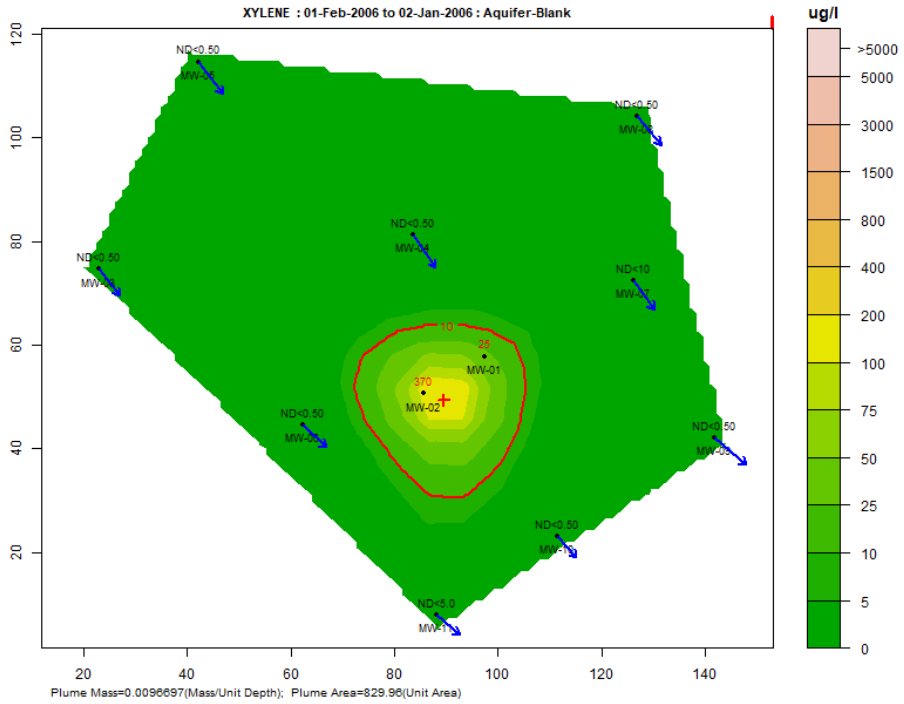

Figure 8: GWSDAT Spatial plot with plume diagnostics.

6.3 Plume Diagnostics

As described in the previous section, clicking the *Plume Diagnostic* checkbox automatically delineates a plume concentration boundary and displays the plume diagnostic quantities underneath the spatial plot, see example in Figure 8. The *Plume Mass, Area & Conc.* and *Estimate Boundary* tabs located to the right of the *Spatial Image* tab, see Figure 7, provide options for further evaluating plume stability.

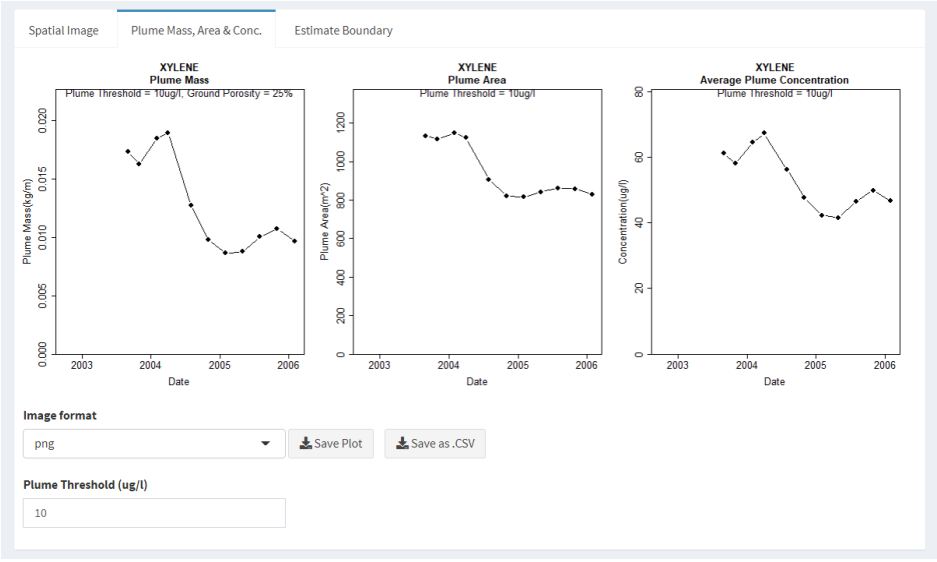

Figure 9: Time series of estimated Plume Mass, Area and Average concentration.

- **Plume Mass, Area & Conc**: This function attempts to calculate, for each time slice, the plume metrics of mass, area and average concentration as described in Appendix 7.2. A time series plot of

these quantities, for each time slice where a closed delineated plume can be formed, is displayed, see example in Figure 9. The user can also select to export the complete time series of plume diagnostics data to a Microsoft Excel csv file. The *Plume Threshold* value (in ug/l), for the currently selected solute, can be adjusted from the numeric input control located at the bottom of this display. The individual *Plume Threshold* values for the complete set of solutes can be set in the *Options* menu, see Section 6.11. Note: Plume metrics are calculated based on the units of length entered in the CoordUnits field in the Well Coordinates table (Section 4.2). If no units are entered then relative changes in plume parameters are plotted in dimensionless units. Note also that plume mass is calculated per unit of plume depth (e.g. kg/m in Figure 9). To estimate total plume mass the user must multiply this value by the estimated plume thickness (using same units as those entered in the CoordUnits field). By default the effective (interconnected) porosity of the subsurface is assumed to be 25%. See Section 6.11 for details on how to modify this value. Additional information on how plume parameters are calculated can be found in Appendix 7.2.

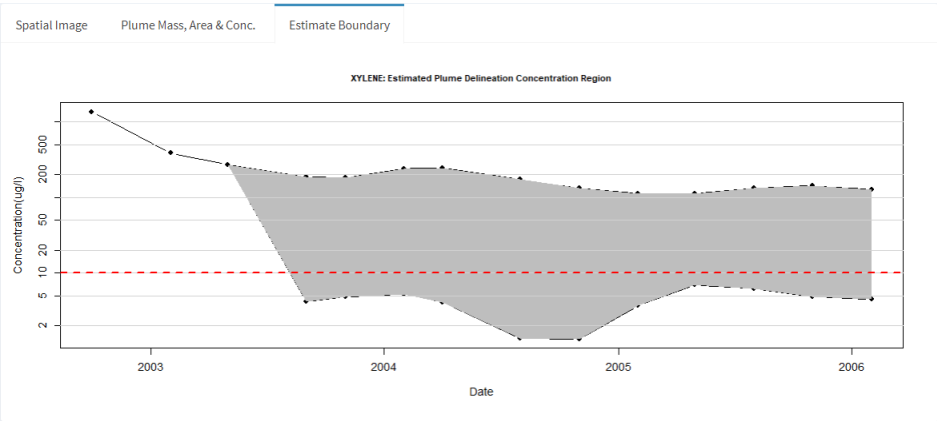

Figure 10: Example Estimate Plume Boundary plot. The grey shaded area defines a region where plume metrics (mass, area, average concentration) can be estimated

- **Estimate (Plume) Boundary**: This function provides a graphic to assist the user in selecting an appropriate solute concentration value that will yield a closed plume, which is an essential requirement for the estimation of plume metrics (i.e. mass, area and average concentration). The plume boundary contour is determined to be closed when it does not extend outside the area bounded by monitoring data. The function works by iterating through each time step of the spatiotemporal model and plotting minimum and maximum solute concentrations at the edge of the area bounded by the monitoring data (see Figure 10). The grey shaded area between the minimum and maximum lines defines the region where plume metrics can be calculated. The user-defined value of the plume threshold concentration value is displayed as a horizontal red dashed line. The user should select a solute concentration value that is just above the minimum line, for the time period within which calculation of plume parameter is required. It is clear from Figure 10 that a value of 10 ug/l works well in this example. The selection of a higher solute concentration value within the grey area, while still producing a closed plume, would exclude part of the plume from the plume mass/ area analysis. The selection of a lower solute concentration value (e.g. 2 ug/l) would, in this example, greatly reduce the time period over which the plume was closed.

Note: The plume diagnostics functions provide insight into relative changes in plume mass, area and average concentration for a given site and monitoring well network. Plume mass estimates are provided in units of kg/m or kg/ft to reflect this uncertainty. Absolute values of plume mass cannot be calculated because the model does not define the vertical concentration distribution inside the plume. The assumptions made in plume mass estimation are that:

- Vertical variation in plume concentration is averaged across monitoring well screen intervals.
- The monitoring well network samples the full depth and extent of the plume and is also representative of concentration variation across the site.

The assumptions are similar to those presented in the paper by Ricker (2008).

6.4 Well Redundancy Analysis

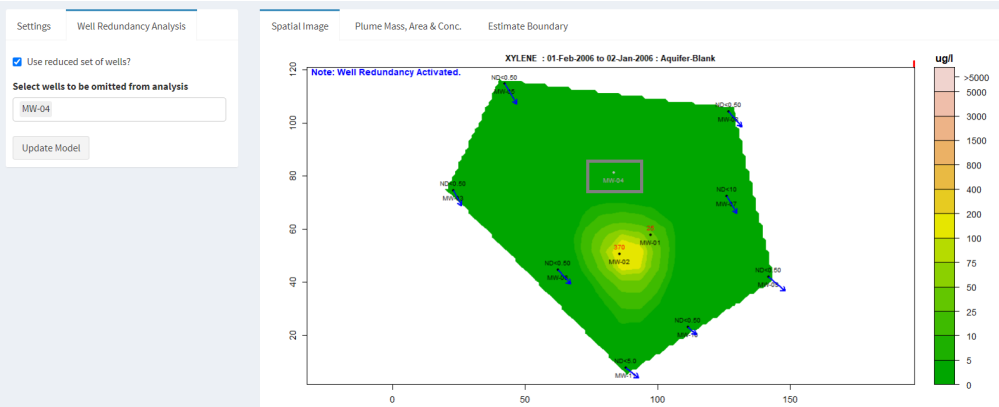

Figure 11: Well redundancy analysis for Spatial plot using the basic example data set having omitted monitoring well MW-04 and using Xylene as the current selection of solute.

The *Well Redundancy Analysis* tab, located next to the *Settings* tab (top left of Figure 7), is a feature first introduced in GWSDAT version 3.1. It allows the user to very conveniently drop a well or a combination of wells from the analysis and investigate the resultant impact. The primary intent of this tool is to understand which wells may have the most influence and also provide supporting evidence that the conclusions of the analysis would not be significantly different with the omission of some certain wells. Figure 11 displays an example Well Redundancy Analysis for the Spatial plot. The listbox *Select wells to be omitted from analysis* allows the user to select any combination of wells from the complete list of wells.

Note: Once the candidate set of wells have been selected the user must manually update the models by clicking the *Update Model* button. This step applies the same spatiotemporal model to each solute as described in Section 7.1 but with a reduced input data set. The calculation of the underlying reduced well data set models can be computationally expensive. Hence the manual decision to update the model when desired, rather than automatically updating the model once any change has been detected in the listbox.

The *Use reduced set of wells?* checkbox enables the user to toggle between the full model and the model using the reduced well data set. The updated predictions of spatiotemporal concentration smoother are plotted and the omitted wells 'greyed' out. The groundwater elevation and flow strength and direction estimates are also updated to the reflect the omission of selected wells. A title is added to the top left of the Spatial plot to notify users that *Well Redundancy Analysis* has been activated.

The plume diagnostic metric quantities as described in Section 6.3 are calculated for both the full (coloured black) and reduced (coloured green) well data set spatiotemporal solute concentration smoother models. Both sets of estimates for plume mass, area and average concentration are overlaid on the same plot, as shown in Figure 12. This allows for a direct side by side comparison. The *Estimate (Plume) Boundary* functionality is also updated to reflect the reduced well data set and a title added to remind the user of this fact.

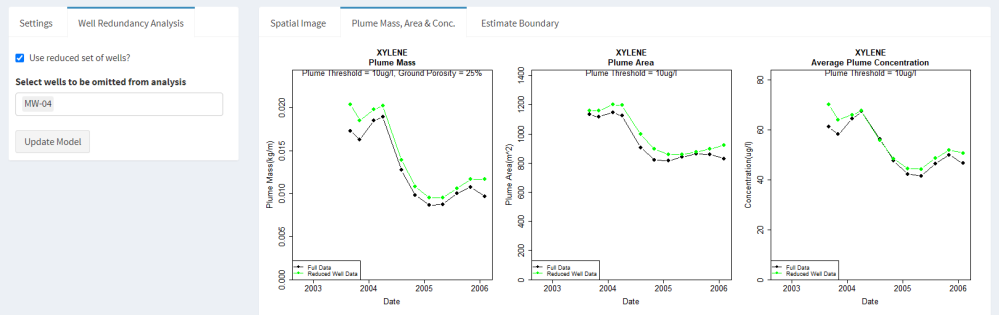

Figure 12: Well redundancy analysis on estimated plume metrics using the basic example data set having omitted monitoring well MW-04 and using Xylene as the current selection of solute. Note: full well data set is color-coded black; reduced well data set is color-coded green.

6.5 Time Series

The Time Series plot enables the user to investigate time series historical trends of solute concentrations in individual wells. Figure 13 displays an example GWSDAT Time Series plot of' 'Benzene' in well 'MW-01' using an illustrative example data set. The actual sampled concentration values are plotted against sampling date and are represented as black solid points. Orange points represent the substituted non-detect values according to the selection chosen in Section 5.2. Red points represent the NAPL substituted solute concentration values.

To switch between different solutes and monitoring wells, simply select from the *Substance* and *Select Monitoring Well* listboxes. The *Solute Conc. Unit* radiogroup allows the units to be changed between one of 'ng/l', 'ug/l' or 'mg/l'. The *Display Threshold* checkbox allows the user to overlay threshold value used to colour code the Trend and Threshold Indicator plot - see Section 6.6.

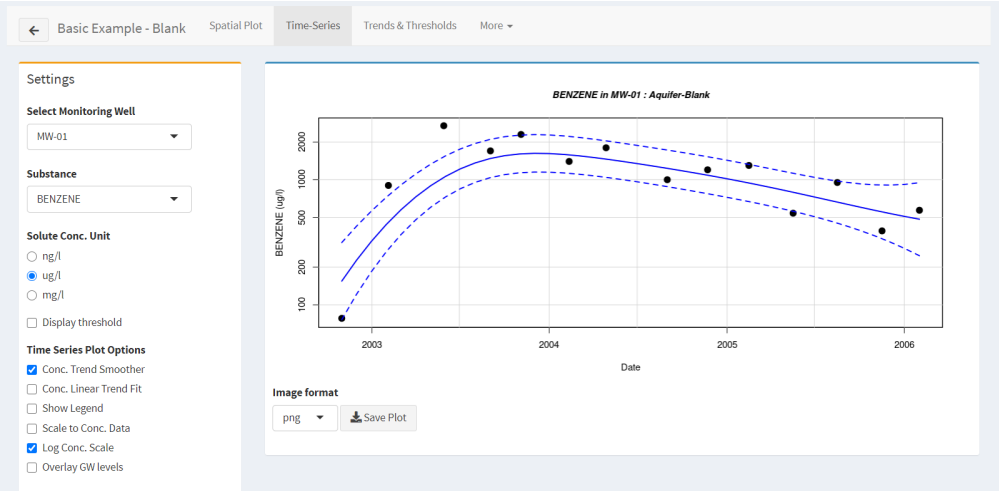

Figure 13: GWSDAT Time Series Plot

The *Time Series Plot Options* checkbox control includes:

- **Conc. Trend Smoother**: This displays the estimated time series trend in solute concentration using a nonparametric smoother (see Figure 13. The solid blue line displays the estimate of the mean trend level at a particular point in time. The upper and lower dashed blue lines depict a 95% confidence interval around this estimate. This is interpreted as *'one is 95% confident that the actual mean trend level lies within this region'*. The smaller the 95% confidence interval, the more confidence one has in the estimated time series trend. Areas of the trend smoother fit in which the 95% confidence intervals are very large (i.e. very low confidence in the trend smoother fit) are coloured grey instead of blue

and are disregarded from the 'Trend' and 'Threshold - Statistical' matrix plot calculations, see Section 6.6. The advantage of this nonparametric method is that the trend estimate is not constrained to be monotonic, i.e. the trend can change direction. More details of this nonparametric smoothing algorithm are given in Appendix 7.4

- **Conc. Linear Trend**: This displays a traditional linear time series trend estimate (green solid line) together with 95% confidence intervals (green dashed lines) to the log of historical solute concentrations values. This is equivalent to fitting an exponential decay/growth model on a linear scale. The statistical significance of this trend is assessed by means of the well established Mann-Kendall trend test Mann (1945). The Mann-Kendall p-value and the estimated solute concentration half-life is displayed immediately below the main title of the *Time Series* plot. Users should be aware that individual well half-life values should not be used to estimate the plume half-life.

If the Mann-Kendall p-value is below 0.05, then the estimated trend is deemed statistically significantly different from 0, i.e. there is indeed trend present in the data. A p-value above 0.05 should be interpreted as there is no evidence to suggest that trend is present.

- **Show Legend**: This controls whether to display a legend in the top right hand side of the plot giving a key of the plotting symbols.
- **Scale to Conc. Data**: By default the *Time Series* plot x-axis is scaled such that it spans the sampling dates of all data. The y-axis is scaled to span the current data concentrations and the user-specified trend threshold limit, see Section 6.6. By checking this control the x and y axes are scaled to the span of the current combination of well and solute concentration data only.
- **Log Scale**: Controls whether to use a logarithmic or linear scale for the y-axis, i.e. solute concentration values.
- **Overlay GW levels**: Allows the user to overlay the corresponding groundwater level measurements on the time series plot. The scale is read from the right hand axis. This function is useful for assessing correlations between groundwater levels and solute concentrations.
- **Overlay NAPL Thickness**: Allows the user to overlay the corresponding NAPL thickness level measurements on the time series plot. The scale is read from the right hand axis. This function is useful for assessing correlations between NAPL thickness and groundwater levels.

6.6 Trends & Thresholds

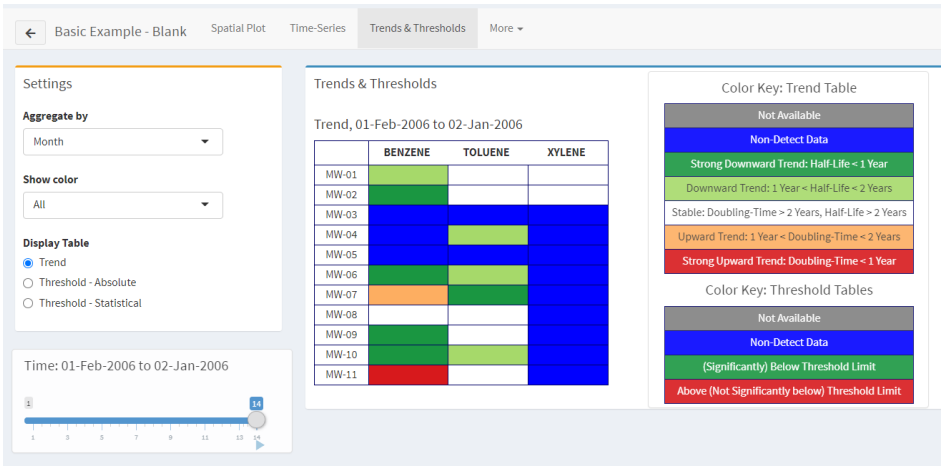

Figure 14: GWSDAT Trends and Thresholds plot.

The *Trends and Thresholds* plot (a.k.a. Traffic Light Plot) is a summary of the strength and direction of the current trend or exceedance relative to a user defined threshold in solute concentrations at a particular

model output interval, see Figure 14. It uses the fitted nonparametric time series trend smoother described in Section 6.5. The rows correspond to each well and the columns correspond to the different solutes. The options are as follows:

- **Display Table**: This drop-down listbox allows the user to select from the following options:
  - **Trend**: This reports the concentration trend for each solute in every well within the selected model output interval. The *Trends Thresholds* plot looks at the instantaneous gradient of the trend smoother (solid blue line) where it crosses the end of the current model output interval in the *Time Series* Plot, see Figure 13. The cells of the *Trends Thresholds* plot are coloured to indicate the strength and direction of the current trend. White cells indicate a generally flat trend where the solute concentration is estimated to no more than double or half in the next two years. Light red and light green indicate that solute concentrations will no more than double or half in the next year, respectively. Dark red and dark green indicate stronger upward and downward trends, respectively. In the event that the trend cannot be calculated, e.g. no data or our confidence in the trend smoother estimate is poor then the corresponding cell is coloured grey. Blue cells represent non-detect data. As an example consider Figures 13 and 14. It can be seen that the trend at the end of the current model output interval (01-Feb-2006) for 'Benzene' at monitoring well 'MW-01' is decreasing. The corresponding cell in Figure 14 (top left) has been coloured light green to illustrate this.
  - **Threshold - Absolute**: This assesses if the **observed** solute concentration values for all well and solute combinations are below a user-specified threshold value (default value of 500 ug/l) within any given model output interval. The threshold value is displayed as a horizontal dashed red line in the *Time Series* Plot, see Figure 13 when the *Display Threshold* checkbox has been selected, see Section 6.5. The 'Threshold - Absolute' option compares the observed concentration values with the threshold value. If any observed concentration values within a model output interval are above the threshold value then the corresponding cell is coloured dark red. If the concentration values within a model output interval are all below the threshold value then the corresponding cell is coloured dark green. In the event that no data exists then the cell is coloured grey. If the current concentration value is classified as non-detect, then the corresponding cell is coloured blue. See Section 6.11 for details on how to edit the threshold values for each solute.
  - **Threshold - Statistical**: This assesses if current solute concentration levels for all well and solute combinations are below a user-specified threshold value **with a statistical degree of confidence**. Again the threshold value is displayed as a horizontal dashed red line in the *Time Series* Plot, see Figure 13 when the *Display Threshold* checkbox has been selected, see Section 6.5.The 'Threshold - Statistical' option looks at the intersection of the end of the current model output interval (vertical grey line) and the trend smoother (solid blue line). If the upper 95% confidence interval (upper dashed blue line) is below the user-specified threshold value, the cell is coloured dark green. If the upper 95% confidence interval is not below the threshold value, the corresponding cell is coloured dark red. In the event that this cannot be calculated, e.g. no data or our confidence in the trend smoother estimate is poor then the cell is coloured grey. If the current concentration value is classified as non-detect, then the corresponding cell is coloured blue. See Section 6.11 for details on how to edit the threshold values for each solute.
- **Show colour**: this drop-down listbox allows the user to filter the *Trends and Thresholds* plot according to the different colours. For example, if the user selects red then the plot will only display the corresponding rows and columns which contain a red entry. This function is particularly useful when there exists a large number of wells and/or solutes.
- **Aggregate by**: provides a drop-down menu which allows the temporal plotting resolution to be altered (Day, Month, Quarter, Year). This functionality is identical to the 'Aggregate by' option detailed in Section 6.2.
- **Colour Key**: A graphic displaying the colour key explained above for the *Trends and Thresholds* plot. Note that this graphic can be moved to any convenient position by clicking and dragging with the

mouse.

- **Temporal Slider**: Use the slider (see bottom left of Figure 14 to explore trends and thresholds across different time slices of the data. Note that this slider box can be moved to any convenient position by clicking and dragging with the mouse. The 'Play' symbol (forward-arrow) in the bottom right hand corner of the slider activates a 'movie', which can be paused by pressing the button again.

6.7 Well Report

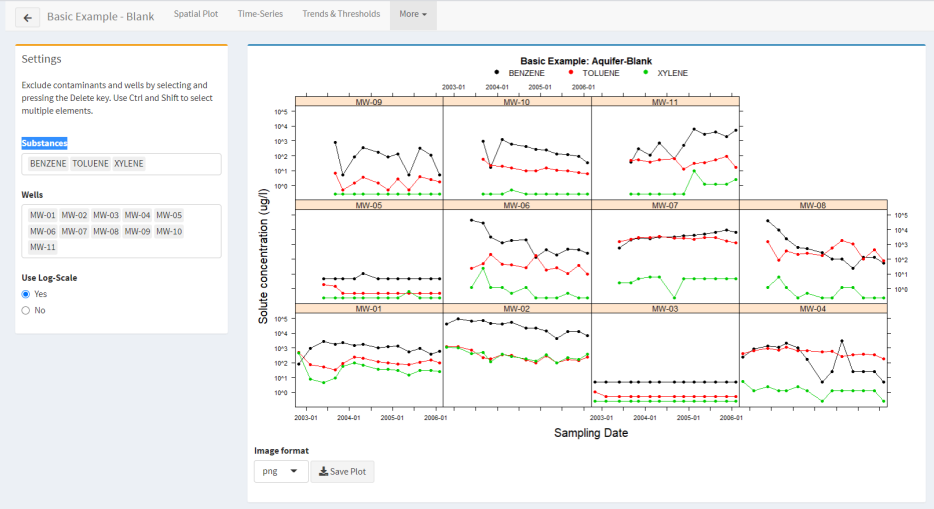

Figure 15: GWSDAT Well Report plot.

This selection, which can be found under the tab named 'More', generates a matrix of graphs displaying time series solute concentration values on a well by well basis. In contrast to the *Time Series* plot (Section 6.5), it is possible to overlay different solute concentration values within the same graph. Figure 15 is an example 'Well Reporting' output. The colour key at the top identifies each solute and the name of each well is displayed in a banner at the top of each of the individual time series graphs. The 'Well Reporting' output provides a very concise method of visualising a lot of data.

The choice of which solutes and wells to include, together with the choice of whether to use a log-scale for the solute concentration values, is selected by the user from the controls on the left hand side of the plot.

If only one solute is selected, then the plotting behaviour is modified such that the detect and non-detect data points are coloured black and orange, respectively. Furthermore, if the 'Conc. Trend Smoother' is checked in Figure 13 then the corresponding trend smoother with 95% confidence intervals are overlaid as thin black lines onto each graph.

6.8 Spatiotemporal Predictions

This selection, located under the 'More' tab, is a visualisation to help assess the goodness of fit of spatiotemporal solute concentration smoother (see Section 6.2 and 7.1) to the observed concentration data. A matrix of graphs displaying time series solute concentration values are generated on a well by well basis and the predictions of spatiotemporal solute concentration smoother are overlaid as solid grey lines, see example in Figure 16. The choice of which solute and wells to include, together with the choice of units and whether to use a log-scale for the solute concentration values, is selected by the user from the controls on the left hand side of the plot.

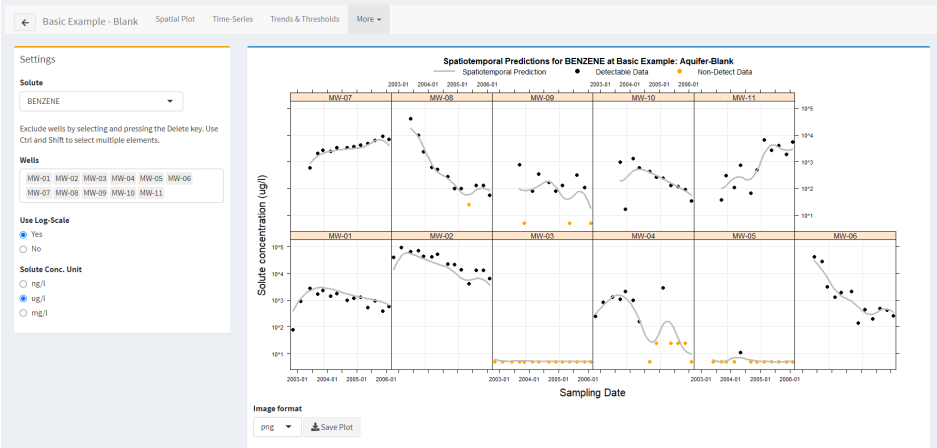

Figure 16: GWSDAT Spatiotemporal Predictions plot.

6.9 Customise Colour Key

This selection, located under the 'More' tab, is a function introduced in version 3.1 to customise the colour key in the *Spatial* plot (see Figure 7 and Section 6.2). Using the data editor, Figure 17, the colour key threshold values can be edited for each solute. Note the save button, located in the bottom right, needs to be clicked in order for the colour key to be updated.

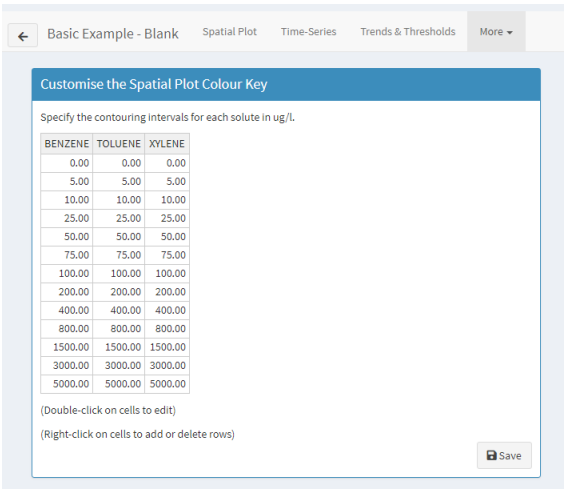

Figure 17: Customising the Spatial plot Colour Key.

6.10 Save Session

This selection, located under the 'More' tab, allows the user to save and download the current analysis session to a '.rds' file format. This is particularly useful for large data sets which may take a long time to fit the underlying models. The current display settings, e.g. Colour Scheme, Plot Options, Current Solute, are saved. This '.rds' file can be forwarded to another user and viewed once again exactly as it was saved.

If using the Excel Add-in data input interface then the *Restore Session* option (see Section 5.1) can be used to open a previously saved '.rds' session. Alternatively, use the *Load Data* function, see Section 6.1.

6.11 Options

This selection, located under the 'More' tab, allows the user to specify the following settings:

- **Concentration Thresholds**: Edit and set concentration threshold values in ug/l for each solute. This is used for colour coding in the Trends and Thresholds visualisation, see Section 6.6.
- **Plume Diagnostic Settings**: Edit and set plume concentration threshold values in ug/l for each solute. This is used for plume diagnostics, see Section 6.3. Also set the ground porosity to a value ranging from 0-100% (default=25%). See Section 7.2 for details on how this is used.
- **Image Export**: Allows the user to change various image export resolution settings, including *jpeg % quality* and pixel dimensions.
- **Model Settings**: This option controls the resolution of the spatiotemporal solute concentration smoother (see Appendix 7.1). In an identical fashion to the Excel Add-in interface (Section 5.2) the user can select between either a default resolution or a higher resolution model fit. In most instances there will be only small differences in the modelling results between the two settings. However, in some rare circumstances with complex data sets, it may well be necessary to use the higher resolution setting. A default resolution model corresponds to a model *Number of segments*=6 and a higher resolution model corresponds to 8. The user can select up to a value of 12. Please note for each unit increase *Number of segments* the model will take 3-4 times longer to fit.

7 Appendices

7.1 Spatiotemporal Solute Concentration Smoother

The spatiotemporal solute concentration smoother is estimated using a non parametric regression technique known as Penalised Splines (P-Splines). It is beyond the scope of this document to give a full and detailed explanation of this technique here. However, the following outlines some of the most important aspects for the purposes of GWSDAT. For a more detailed explanation the reader is referred to P. Eilers and Marx (1992) and P. H. C. Eilers, Rijnmond, and Marx (1996).

Let y_i be the solute concentration at $\boldsymbol{x_i} = (x_{i1}, x_{i2}, x_{i3})$ where x_{i1} and x_{i2} stand for the spatial coordinates of the well and x_{i3} represents the corresponding time point for the *i-th* observation with $i = 1, \ldots, n$. We start by modelling the solute concentration as

$$y_i = \sum_{j=1}^{m} b_j(\boldsymbol{x_i})\alpha_j + \epsilon_i \tag{1}$$

where the b_j, $j = 1, \ldots, m$ are m functions (known as *basis functions*) conveniently chosen to achieve smoothness (generally a particular kind of polynomial of order 3). The first term in equation (1) is a linear combination of the basis functions b_j, each evaluated at $\boldsymbol{x_i}$, and aims at capturing the deterministic part of the *i-th* observation, generally known as 'signal'; the second term, ϵ_i, accounts for the variability in the measurement due to randomness and is usually termed as 'noise'. The behaviour of ϵ_i is described in terms of a convenient probabilistic model; such a model guarantees that the value of ϵ_i fluctuates around zero conveying the idea that we do not expect to make any systematic error in the measurement. This model also comprises the notion that the expected spread of ϵ_i is given by σ^2, %a non-negative parameter σ; its squared %value is known as the *variance* of the random component ϵ_i. By using the matrix notation

$$\boldsymbol{B}(\boldsymbol{x}) = \begin{pmatrix} b_1(x_1) & \cdots & b_j(x_1) & \cdots & b_m(x_1) \\ b_1(x_2) & \cdots & b_j(x_2) & \cdots & b_m(x_2) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_1(x_i) & \cdots & b_j(x_i) & \cdots & b_m(x_i) \\ \vdots & \cdots & \vdots & \cdots & \vdots \\ b_1(x_n) & \cdots & b_j(x_n) & \cdots & b_m(x_n) \end{pmatrix}$$

equation (1) can be written in a more compact fashion as $\boldsymbol{y} = \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha} + \boldsymbol{\epsilon}$. Because, as mentioned earlier, we expect the ϵ_i's to oscillate around zero, a sensible choice for the regression parameters $\boldsymbol{\alpha}$ is the one that minimises the norm of the vector $\boldsymbol{\epsilon}$ defined as $\mathrm{S}(\boldsymbol{\alpha}) = \|\boldsymbol{\epsilon}\|^2 = \|\boldsymbol{y} - \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha}\|^2$. A large value of basis functions is generally chosen to allow the model to capture most of the signal. The downside of this approach is that it tends also to overfit, that is to fit the noise in the observations, with the consequent loss of smoothness. To overcome this hurdle, the objective function %to be optimised $\mathrm{S}(\boldsymbol{\alpha})$ is modified with the addition of a term that penalises the lack of smoothness of the fit.

The objective function now takes the form $\mathrm{S}(\boldsymbol{\alpha}) = \|\boldsymbol{y} - \boldsymbol{B}(\boldsymbol{x})\boldsymbol{\alpha}\|^2 + \lambda\|D\alpha\|^2$ where λ is a non-negative smoothing parameter and D is the $(m-2) \times m$ matrix

$$\boldsymbol{D} = \begin{pmatrix} 1 & -2 & 1 & 0 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 1 & -2 & 1 & 0 & \cdots & 0 & 0 & 0 \\ 0 & 0 & 1 & -2 & 1 & \cdots & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots & & \\ 0 & 0 & 0 & 0 & 0 & \cdots & 1 & -2 & 1 \end{pmatrix}$$

The additional term in the objective function

$$\|D\alpha\|^2 = (\alpha_1 - 2\alpha_2 + \alpha_3)^2 + \ldots + (\alpha_{m-2} - 2\alpha_{m-1} + \alpha_m)^2$$

controls the smoothness of the fit by applying penalties over adjacent coefficients. By minimising the new objective function for a given value of λ, we obtain the least squares estimator of the parameters

$$\hat{\boldsymbol{\alpha}} = \left(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D}\right)^{-1} \boldsymbol{B}'\boldsymbol{y}.$$

Consequently, the fitted values are given by:

$$\hat{\boldsymbol{y}} = \boldsymbol{B}\hat{\boldsymbol{\alpha}} = \boldsymbol{B}\left(\boldsymbol{B}'\boldsymbol{B} + \lambda \boldsymbol{D}'\boldsymbol{D}\right)^{-1} \boldsymbol{B}'\boldsymbol{y} = \boldsymbol{H}\boldsymbol{y}$$

When $\lambda = 0$, the expression for the estimator of the parameters $\{\ \hat{\boldsymbol{\alpha}}\ \}$ boils down to the classical solution in linear models theory. As $\boldsymbol{\lambda} \to \boldsymbol{\infty}$, the fitted function tends to a linear function. The Figure below shows the effect of penalisation: it forces the coefficients to yield a smooth pattern. The fitting process of a function using B-Splines is pictured with and without penalisation, together with the basis functions (the columns of the $\boldsymbol{B}$ matrix). The left plot results from not penalising ($\boldsymbol{\lambda = 0}$) the term in the objective function that accounts for the smoothness; it can be noticed that it yields a rather wiggly regression function. In the right plot, a suitable choice for $\boldsymbol{\lambda}$ constrains the optimisation method to find values for the coefficients $\hat{\boldsymbol{\alpha}}$ which result in a smoother regression curve.

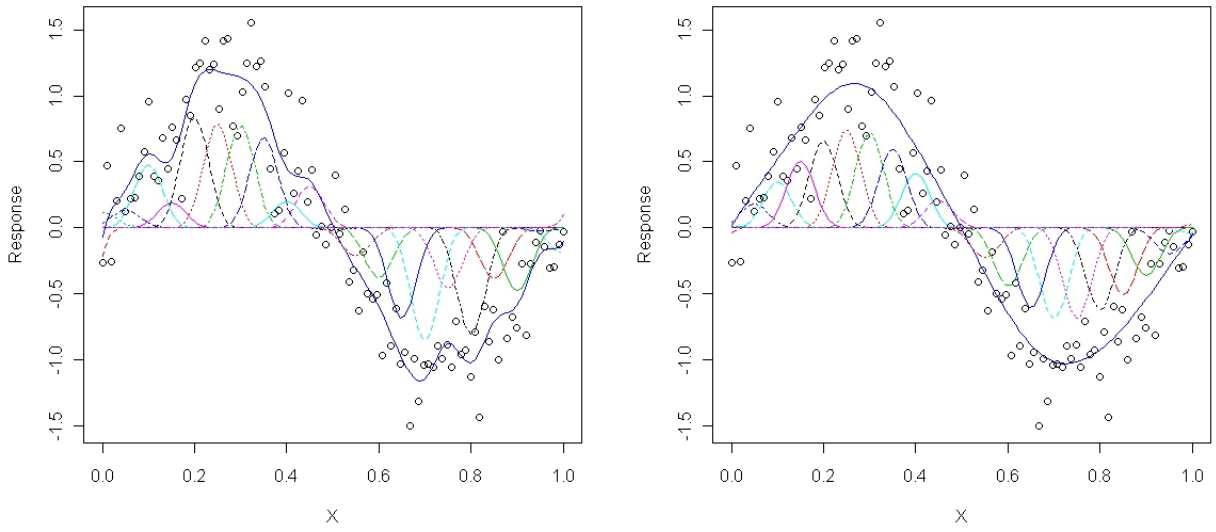

fig.cap: Curve based on 20 nodes in the basis, without penalisation (left), with penalisation (right).

Prior to fitting the regression coefficients $\boldsymbol{\alpha}$ the observed solute concentration values are natural log transformed. This avoids the possibility of predicting negative concentration values and also helps the model cope with data which often spans several orders of magnitude. Furthermore, the uncertainty in the measured concentrations can reasonably be expected to be proportional to the magnitude of the value, e.g. the uncertainty around a measured value of 10ug/l would be expected to be very much less than the uncertainty surrounding a measured value of 10000ug/l. The natural log transformation stabilises the variance.

The choice of the penalisation parameter $\boldsymbol{\lambda}$ is a crucial matter as a too small value would result in 'overfitting' (tracking the noise) whereas an extremely large value would lead to 'underfitting' (producing a flat estimated function as a result of loss of signal). Several criteria have been proposed, such as those described by C. Hurvich and Simonoff (1998) and Simon. N. Wood (2006), but we tackled the issue by a *Bayesian* approach; see Denison et al. (2002), Raftery, Madigan, and Hoeting (1997) and S. N. Wood (2011).

Under this paradigm, $\boldsymbol{\lambda}$ is not considered to be a fixed unknown quantity to be estimated but rather a random variable whose value may vary within a given range. This behaviour is described in probabilistic terms which assign a measure of confidence or *probability* to each of the values $\boldsymbol{\lambda}$ may take on.

The Bayesian framework allows to compute the probability that the random variable $\boldsymbol{\lambda}$ may take a particular value, conditional on the fact that $\boldsymbol{y}$ has already been observed. This probability, indicated as $\boldsymbol{f(\lambda|y)}$, is known as the *posterior distribution* of $\boldsymbol{\lambda}$.

Bayes' rule states that $\boldsymbol{f(\lambda|y) \propto f(y|\lambda)f(\lambda)}$ where $\boldsymbol{\propto}$ stands for "proportional to". $\boldsymbol{f(y|\lambda)}$ is known as the *likelihood function* and expresses the conditional probability of observing data $\boldsymbol{y}$, given that the true value of the parameter is $\boldsymbol{\lambda}$; $\boldsymbol{f(\lambda)}$ is known as the *prior distribution* of the random variable $\boldsymbol{\lambda}$ and comprises our prior beliefs on its uncertainty.

The optimal value of $\boldsymbol{\lambda}$ is the one that maximises the posterior distribution and is computed using numerical methods.

7.2 Plume Diagnostics

GWSDAT calculates plume diagnostic quantities from the predictions of the `Spatiotemporal Solute Concentration Smoother`. In common to Aziz et al. (2003) and Ricker (2008), numerical methods are employed to integrate out the plume diagnostic quantities. For a given model time step a fine spatial mesh grid of predictions is generated. The plume boundary region $\boldsymbol{D}$, for a given plume threshold concentration value, is calculated using the R function *contourLines* which is included in the base distribution of the R programming language (R Development Core Team (2008)). The plume area, $\boldsymbol{A}$, is defined as

$$\boldsymbol{A = \iint_D dxdy} \tag{2}$$

where where $\boldsymbol{x}$ and $\boldsymbol{y}$ are the spatial coordinates and is calculated numerically using the *areapl* function from the R package splancs Rowlingson et al. (2021). The average plume concentration is defined as

$$\boldsymbol{\frac{1}{A} \iint_D \hat{s}_t\,(x, y)\,dxdy} \tag{3}$$

where $\boldsymbol{\hat{s}_t\,(x, y)}$ represents the predictions of the spatiotemporal solute concentration smoother evaluated at time $\boldsymbol{t}$. This integral (and all subsequent integrals in this section) is calculated numerically using a method described in Oloufa (1991). A Delaunay triangulation is performed using the R package 'deldir', Turner (2021), on the spatial mesh grid of predictions within the plume boundary, $\boldsymbol{D}$.
The integral is numerically approximated by summing up the individual volumes under each prism formed.

Plume mass is calculated from the scaled product of plume area and average concentration. The scaling factor encompasses the user specified value of ground porosity (see `Plume Diagnostics`) and appropriate scaling values for mapping together the volumetric concentration units (e.g. ug/l) with the length scale (see **CoordUnits** in `Well Coordinates Table`) of the Well coordinates (e.g. metres or feet). The plume mass is calculated on a mass per unit aquifer depth basis (e.g. kg/m). To calculate the total plume mass the user must multiply this value by the aquifer depth.

The plume center of mass $\boldsymbol{(x, y)}$ is defined as the mean location of the concentration distribution within the plume boundary region $\boldsymbol{D}$. The x-coordinate of the plume Centre of Mass is evaluated by numerical calculation of

$$\boldsymbol{X_c = \frac{\iint_D x\hat{s}_t\,(x, y)\,dxdy}{\iint_D \hat{s}_t\,(x, y)\,dxdy}} \tag{4}$$

where x and y are the spatial coordinates and $\hat{s}_t(x, y)$ represents the predictions of the spatiotemporal solute concentration smoother evaluated at time t. In a completely analogous manner the y-coordinate of the plume Centre of Mass is evaluated as follows:

$$Y_c = \frac{\iint_D y\hat{s}_t(x, y)\, dxdy}{\iint_D \hat{s}_t(x, y)\, dxdy} \tag{5}$$

In the event that multiple plumes are detected then the above quantities are calculated for each individual plume and aggregated together. The individual plume areas and masses are summed to calculate the total over all plumes. The aggregate average plume concentration and aggregate plume centre of mass is calculated by taking a weighted average of the individual quantities.

7.3 Groundwater Flow Calculation

For a given model output interval the Groundwater (GW) flow strength and direction are estimated using available GW level and well coordinates data. The model is based on the simple premise that local GW flow will follow the local direction of steepest descent (hydraulic gradient).

For a given well, a linear plane is fitted to the local GW level data:

$$L_i = a + bx_i + cy_i + \epsilon_i \tag{6}$$

where L_i represents the GW level at location (x_i, y_i). Local data is defined as the neighbouring wells as given by a Delauney triangulation (http://en.wikipedia.org/wiki/Delaunay_triangulation, Ahuja and Schacter (1983)) of the monitoring well locations. The gradient of this linear surface in both x and y directions is given by the coefficients b and c. Estimated direction of flow is given by:

$$\theta = \tan^{-1}\left(\frac{c}{b}\right) \tag{7}$$

and the relative hydraulic gradient (a measure of relative flow velocity) is given by

$$R = \sqrt{b^2 + c^2} \tag{8}$$

For any given model output interval this algorithm is applied to each and every well where a GW level has been recorded.

7.4 Time Series Plot Smoother

The time series plot smoother is fitted using a nonparametric method called local linear regression. This involves solving locally the least squares problem:

$$\min_{\alpha,\beta} \sum_{i}^{n} \{y_i - \alpha - \beta(x_i - x)\}^2 w(x_i - x; h) \tag{9}$$

where $w(x_i - x; h)$ is called the kernel function. A normally-distributed probability density function with standard deviation h is used as the kernel. h is also called smoothing parameter that controls the width of the kernel function, and hence the degree of smoothing applied to the data (the higher the value of h, the smoother the estimates). Within GWSDAT, local linear regression is deployed using the R package 'sm' (Bowman and Azzalini (1997)) and the bandwidth is selected using the method published in C. M. Hurvich, Simonoff, and Tsai (1998).

7.5 Converting a CAD drawing to a Shapefile

System requirements: ArcGIS comprising ArcMap, ArcEditor, ArcCatalog

1. Open ArcCatalog from the Start Menu ('Start' -> 'All Programs' -> 'ArcGIS' -> 'ArcCatalog').
2. In ArcCatalog navigate to ArcMap (globe with magnifying glass icon).
3. When ArcMap opens a screen will pop-up. Select 'A New Empty Map' then click 'OK'.
4. Go to 'File' -> 'Add Data' (positive sign with yellow triangle underneath) -> Select site CAD drawing saved as a '.dxf' file -> 'Add'.
5. Click on the '+' symbol to expand the sub-layers of the dxf file (e.g. 'Polyline', 'Polygon', 'Multipatch', 'Point').
6. Right click on required layer (e.g. Polyline or an edited & exported shapefile) to open the drop down menu.
7. On drop down menu select 'Data' -> 'Export Data'.
8. On 'Export Data' pop-up menu choose 'Select All Features' + 'This Layers Source Data' and select the folder you wish to save the shapefile into, then click 'OK'.
9. Click 'Yes' to add the exported data as a new layer.
10. Repeat steps 6-9 to convert all the layers required to produce the base-map in GWSDAT into shapefiles.
11. Add the shapefiles into GWSDAT (see `GIS ShapeFiles Table`) one by one to produce the complete base-map image.

The next section details how to edit layers in ArcMap after their conversion to shapefiles, prior to upload into GWSDAT (useful for removing gridlines etc)

1. Uncheck the CAD layer used to produce shapefile to remove image from view window.
2. Ensure exported shapefile is selected and visible in view window.
3. Click 'Start Editing' on the 'Editor' toolbar above the map.
4. Use the arrow pointer to select lines and press delete on the keyboard to remove from drawing. (Select 'UnDo' from 'Edit' Toolbar in case of errors).
5. 'Editor' -> 'Stop Editing'. Click 'Yes' to save edits.
6. Repeat data export as detailed in steps 6-9 above and re-save as new shapefile.

References

Ahuja, N., and B. J. Schacter. 1983. *Pattern Models.* John Wiley & Sons, New York.

Aziz, Julia A., C. J. Newell, Meng Ling, Hanadi S. Rifai, and J. R. Gonzales. 2003. "MAROS: A Decision Support System for Optimizing Monitoring Plans." *Ground Water* 41 (3): 355–67.

Bowman, A. W., and A. Azzalini. 1997. *Applied Smoothing Techniques for Data Analysis:the Kernel Approach with s-Plus Illustrations.* Oxford University Press, Oxford.

Denison, D., C. Holmes, B. Mallick, and A. Smith. 2002. *Bayesian Methods for Nonlinear Classification & Regression.* John Wiley & Sons, New York.

Eilers, Paul H. C., Dcmr Milieudienst Rijnmond, and Brian D. Marx. 1996. "Flexible Smoothing with b-Splines and Penalties." *Statistical Science* 11: 89–121.

Eilers, P., and B. Marx. 1992. *Generalized Linear Models with p-Splines in Advances in GLIM and Statistical Modelling (l.fahrmeir Et Al. Eds.).* Springer, New York.

Hurvich, C. M., J. S. Simonoff, and C.-L. Tsai. 1998. "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion." *Journal of the Royal Statistical Society, Series B* 60: 271–93.

Hurvich, C., and J. Simonoff. 1998. "Smoothing Parameter Selection in Nonparametric Regression Using an Improved Akaike Information Criterion." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 60: 271–93.

Mann, H. B. 1945. "Nonparametric Tests Against Trend." *Econometrica* 13: 245–59.

McLean, M. I., L. Evers, A. W. Bowman, M. Bonte, and W. R. Jones. 2019. "Statistical Modelling of Groundwater Contamination Monitoring Data: A Comparison of Spatial and Spatiotemporal Methods." *Science of The Total Environment* 652: 1339–46. https://www.sciencedirect.com/science/article/pii/S0048969718341275.

Oloufa, A. A. 1991. "Triangulation Applications in Volume Calculation." *Journal of Computing in Civil Engineering* 5 (1): 103–19.

R Development Core Team. 2008. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. http://www.R-project.org.

Raftery, A., D. Madigan, and J. Hoeting. 1997. "Bayesian Model Averaging for Linear Regression Models." *Journal of the American Statistical Association* 92: 179–91.

Ricker, Joseph A. 2008. "A Practical Method to Evaluate Ground Water Contaminant Plume Stability." *Ground Water Monitoring and Remediation* 28 (4): 85–94.

Rowlingson, Barry, Peter Diggle, adapted, packaged for R by Roger Bivand, pcp functions by Giovanni Petris, and goodness of fit by Stephen Eglen. 2021. *Splancs: Spatial and Space-Time Point Pattern Analysis.* http://www.maths.lancs.ac.uk/~rowlings/Splancs/.

Turner, Rolf. 2021. *Deldir: Delaunay Triangulation and Dirichlet (Voronoi) Tesselation.* http://www.math.unb.ca/~rolf/.

Wood, S. N. 2011. "Fast Stable Restricted Maximum Likelihood and Marginal Likelihood Estimation of Semiparametric Generalized Linear Models." *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 73: 3–36.

Wood, Simon. N. 2006. *Generalized Additive Models - an Introduction with r.* Chapman & Hall/CRC.